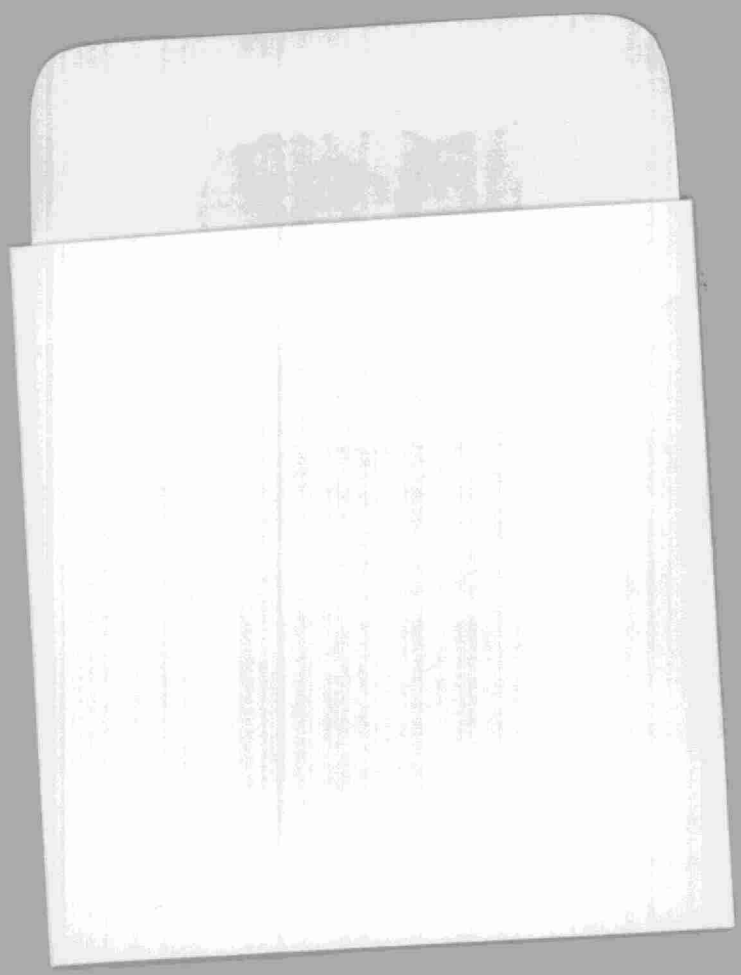
THE

ONTARIO WATER RESOURCES

COMMISSION

Reliability and Confidence in Computing the Dissolved Oxygen Sag

1969



RELIABILITY AND CONFIDENCE

IN COMPUTING

DISSOLVED OXYGEN SAG

by

Dr. Syed S. Ahmed Rizvi
Water Quality Surveys Branch

1969
Ontario Water Resources Commission

RELIABILITY AND CONFIDENCE IN COMPUTING DISSOLVED OXYGEN SAG

Dr. Syed S. Ahmed Rizvi

ABSTRACT

Mathematical modelling of streams is frequently based on a set of dissolved oxygen data collected during intensive field surveys. This paper illustrates, based on statistical techniques, what confidence and reliability one can place on the sag curve drawn from the collected data. The statistical techniques used, thus, give information on the reliability and confidence limits of the mathematical model. Two case studies for different streams illustrate the use of these statistical techniques. The number of water quality samples necessary to improve the confidence limits of the collected data is discussed.

RELIABILITY AND CONFIDENCE IN COMPUTING DISSOLVED OXYGEN SAG

TABLE OF CONTENTS

1. IN TRODUCTION

The capacity of a stream to receive and oxidize sewage or other polluted matter depends to a large extent upon its oxygen resources. The condition of a polluted stream at any time is the result of a balance between these resources and the demand made upon them by the oxygen-using matter carried by the stream. This demand, usually the result of biochemical processes, is, in the absence of new pollution, a progressively decreasing one as one moves downstream. As the resources of the stream are composed in part of a continuous influx of oxygen from the atmosphere, the state of balance which determines the momentary condition of the stream is constantly changing. There are, therefore, two primary phases in the problem; namely, the actual, momentary condition and the direction and extent of the existing changes which indicate the future condition.

This paper presents statistical techniques which can be used in an effort to determine the validity and realiability of sample data of dissolved oxygen (DO) and recommends a minimum practical number of samples required to determine the dissolved oxygen depletion curve of a stream. A case study for stream A is presented to show the reliability and confidence one can place on samples collected over a period of 72 hours. For stream B, two sets of data,

one containing 8 samples, the other 72, all taken over a period of 3 days, are compared to find out if the two samples came from the same population.

2. STATISTICAL TECHNIQUES

Generally, samples are taken at a number of different stations over a time period of 72 hours and a sag curve is drawn through the mean of the samples for each station. The question is what kind of reliability can be placed on the curve and within what confidence limits can the DO data collected be used. In other words, does the mean of the set of observations, $\bar{x}$, closely approximate the population mean μ?

The Central Limit Theorem states that if $\bar{x}$ is the mean of a random sample of size n from any population N, with the mean μ and the variance σ^2, then the sample distribution of $\bar{x}$ is the normal distribution. This justifies approximating the distribution of $\bar{x}$ with a normal distribution with a mean μ and variance $\frac{\sigma^2}{n}$. It is of interest to note that if the common distribution of the random variables, x, is normal, the distribution of $\bar{x}$ is the normal distribution for any n. Most of the time, when the standard deviation σ of a population is unknown, it is estimated from the actual observations and designated s. However,

the sample must be sufficiently large, $n \geq 30$, in order to get a close estimate of the population standard deviation, σ . The precision of s increases as the sample becomes larger.

In general, in order to find a confidence interval for a parameter θ of a given population, we must find two random variables, θ_1 , and θ_2, for which θ_2 can never assume a value less than θ_1, and for which we can assert with a probability of $1 - \alpha$ that they will assume values satisfying the double inequality $\theta_1 < \theta < \theta_2$. It is customary to refer to θ_1 , and θ_2 as the lower and upper confidence limits for θ , to $1 - \alpha$ as the degree of confidence, and to the interval from θ_1 , to θ_2 as the confidence interval for θ .

Referring to the distribution of $\bar{x}$ for random samples from a normal population with a mean μ and variance σ^2, we can assert that with a probability of $1 - \alpha$, the random variable $\frac{\mu - \bar{x}}{\sigma/\sqrt{n}}$ will assume a value between $-z_{\alpha/2}$ and $z_{\alpha/2}$. It can be shown as:

$$-z_{\alpha/2} < \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} < z_{\alpha/2} \qquad \ldots 2.1$$

or

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \ldots 2.2$$

The double inequality 2.2 can only be true or false: either μ is contained between $\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ or it is

not. The value of $z_{\alpha/2}$ can be found in appropriate statistical tables.[1] This double inequality is more apt to be true than false because the value of α is generally taken as $\alpha = .05$, so that there is a probability of 0.95 that $\frac{(\bar{x}-\mu)}{\sigma/\sqrt{n}}$ will assume a value between $\pm z_{\alpha/2}$ or $\pm z_{.025}$. Although one cannot make such a probability statement about μ, one can assert with a probability of $1 - \alpha$ that the random variables $\bar{x} - z_{/2} \frac{\sigma}{\sqrt{n}}$ and $\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ will assume values satisfying 2.2. Therefore, one can say that the probability of a normal variate falling in the range $\pm z = 2 \; p(z)$; while the probability of a variate falling outside the range $\pm z = (1 - 2p(z))$.

The confidence interval given above was designed to estimate the mean of a normal population whose variance, σ^2, is known. When dealing with samples which are large enough ($n \geq 30$) to justify use of the Central Limit Theorem, equation 2.2 is also used to estimate means of other populations with known variances.

The method by which one can construct confidence intervals consists, essentially, of finding an appropriate random variable whose values can be calculated on the basis of the sample values and the parameter, but whose distribution does not depend on

---

[1] Freund, J. E., Mathematical Statistics, Prentice-Hall, p. 366.

the parameter. The random variable, $\frac{(\bar{x} - \mu)}{\sigma / \sqrt{n}}$, was discussed above. Now a similar inequality can be developed for the random variable $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

In order to construct a $(1 - \alpha)$ confidence interval for μ where σ is unknown, one can make use of the fact that for a random sample of size n from a normal population, $\frac{(\bar{x} - \mu)}{s/\sqrt{n}}$ has a t-distribution with $n - 1$ degrees of freedom. The proof for this statement can be found in most books on statistics.[1] Hence, one can assert with a probability of $1 - \alpha$ that this random variable will assume a value between $-t_{\alpha/2,\,(n-1)}$ and $+t_{\alpha/2,(n-1)}$ and for a given sample we assign a degree of confidence of $1 - \alpha$ to

$$-t_{\alpha/2,(n-1)} < \frac{\bar{x} - \mu}{s/\sqrt{n}} < t_{\alpha/2,(n-1)} \qquad \ldots 2.3$$

or

$$\bar{x} - t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} < \mu < t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} + \bar{x} \qquad \ldots 2.4$$

This confidence interval for μ can only be used for random samples from normal populations. For other populations, an approximate confidence interval for μ of a large sample ($n \geq 30$) may be obtained by substituting s for σ in equation 2.2.

---

[1] Freund, J.E., Mathematical Statistics, Prentice-Hall, p. 203.

The above statistic can also be used for comparing means of samples for the purpose of determining whether the observed difference is due to chance only or whether we should suspect some real cause to be responsible and hence consider the difference to be statistically significant.

3. METHOD AND DISCUSSION

A reach of stream A, Figure 1, was selected and thirty-six stations were established along the river. At each site from nine to thirty-six cross-sectional samples were collected. The temperature and dissolved oxygen were measured and recorded at the time of sampling, while BOD_5[1] determinations were made in accordance with the "Standard Methods for the Examination of Water, Sewage and Industrial Wastes" at the OWRC laboratory in Toronto. The stations with the number of samples taken at each section (group of stations) and the mean values of all the measurements across a section are given in Table 1.

To illustrate the use of the above discussed techniques, suppose that measurements of DO values at a certain station may be looked upon as a random sample from a normal population. If, at Station 1, the value of 18 such samples had a mean of 7.97 ppm

---

[1] 5-day, 20°C, Biochemical Oxygen Demand.

(Table 1, line 1) and a standard deviation of 0.66, then with 0.95 confidence, the "true" mean value of DO (ppm) is obtained by substituting these values in equation 2.4. As the value of α is taken as .05, the value for $t_{\alpha/2,(n-1)}$ is taken from statistical tables[1] as $t_{.025,17} = 2.11$. Thus, we get from equation 2.4:

$$7.64 < \mu < 8.30$$

From this, one can assert with a degree of confidence of 0.95 that the interval from 7.64 to 8.30 ppm contains the true mean of DO at this particular station of the stream.

The following parameters were calculated for stream A: mean, standard deviation, standard error of the mean, upper and lower confidence limits. Table 1 shows the results of the calculations along with a minimum and maximum readings for each section. The confidence levels on the true mean were calculated using the t-distribution (equation 2.4) assuming a normally-distributed population.

The mean DO values from Table 1 were plotted and joined smoothly to produce the sag curve. Two smooth curves showing the upper and lower confidence limits were drawn through the values calculated and presented in Table 1 creating an envelope

---

[1] Freund, J.E., Mathematical Statistics, Prentice-Hall, p. 367.

around the mean value curve. The upper and lower curves were based on 95% confidence. This meant that there was a 95% probability that the true population mean lay within this envelope at any point. The curves are shown in Figure 2.

A mathematical model was then developed which reproduced the observed dissolved oxygen sag curve as shown in Figure 2 (solid line). Observing this curve, it is noticed that this mathematical model lies within the 95% confidence envelope at all points except below one large industrial waste source, shown by arrows in Figure 2. This deviation could be explained by two factors: the presence of either a floating, heated waste and/or buoying sludge mats.

On stream B, samples from two adjacent stations were compared one with 72 samples, the other with eight, in order to find out if the two samples came from the same population. All the samples were taken over a period of 72 hours. Table 2 shows the data collected on stream B.

The data from stream B given in Table 2 were used to compare the means of the two samples. The calculated mean, variance and standard deviation along with the results of the F and t tests and the calculated confidence limits are shown in Table 2.

Based on this, it is clear that the two samples did not appear to have come from the same population.

4. SUMMARY AND CONCLUSIONS

Several important results came out of this study. Statistical techniques were presented and developed for a quick and relatively simple method of calculating confidence limits for the mean values of dissolved oxygen (DO) data of a stream. Other results provided more information on selecting a suitable number of samples required in studying the dissolved oxygen sag curve.

From the discussion in the earlier part of this report, it was apparent that if random samples were taken from a normal population or if one assumed a normal population, as in this case study, then one could use the inequality,

$$\bar{x} - t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{\alpha/2,(n-1)} \frac{s}{\sqrt{n}} \quad \ldots 4.1$$

to calculate the confidence limits for the true mean value μ. However, for other populations an approximate confidence interval for μ could be found if $n \geq 30$ where n is the number of samples, by using s instead of σ in equation 2.2 as follows:

$$\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \quad \ldots 4.2$$

These two equations gave a general method of calculating the confidence limits of the mean values for the DO data.

In the case studied for stream A, a dissolved oxygen sag curve with an envelope showing the 95% confidence limits was drawn so that there was a 95% probability that the true mean lay within the upper and lower confidence limits as shown in Figure 2. It is interesting to note that the mathematical model developed from the available data lies within the envelope except at two points.

For stream A, it was also found that the greater the number of samples (n) taken, the greater the reliability that could be placed on the calculated mean value. By taking $n \geq 30$, the assumption of a normal distribution was no longer necessary since equation 4.2 could then be used. However, it should be noted that the standard deviation of the mean varies inversely as the square root of the number of samples because $s_{\bar{x}} = \frac{s}{\sqrt{n}}$. The increase in the reliability is sometimes not worthwhile compared to the effort and cost involved in collecting and analyzing the samples required and calculating the results. As an illustration, a sample of 16 observations is only twice as precise as a sample of 4, so that the gain in precision is small relative to the effort in taking the additional 12 observations.

The case study on stream B gave a further reason for taking a large number of samples. From Table 2, it was clear that two samples though taken over the same time period at the same site apparently did not come from the same population. The smaller sample of eight values did not appear to have originated from the same population as that of the larger sample of 72 values. Since a larger sample indicates mean values closer to the true population mean, a larger sample where possible should be taken.

One general recommendation concerning the number of samples that thirty or more samples should be taken (i.e. $n \geq 30$) at each station. However, it is not always possible to take thirty or more samples due to the cost and other factors involved. As discussed earlier, under these conditions it is necessary to make the assumption that the sample was taken from a population with a normal distribution. An alternative is to place continuous recording meters at selected points along the river. By this method, large samples could be developed and a close estimate of σ could be made. From the continuous records one could also obtain the population distribution. How many of these recording stations within a reach of the stream are required and what distance should be maintained between these meters needs further study.

TABLE I

DISSOLVED OXYGEN CALCULATIONS FOR STATIONS

STREAM "A"

STATION IDENTIFICATION NUMBERS	NUMBER OF SAMPLES (n)	MEAN PPM. ($\bar{X}$)	STANDARD DEVIATION (s)	STANDARD ERROR OF MEAN $s/\sqrt{n}$	MAXIMUM	MINIMUM	FACTOR $\left[t_{\alpha/2,(n-1)}\right]\frac{s}{\sqrt{n}} = (B)$	LOWER CONFIDENCE LEVEL ($\bar{X} - B$)	UPPER CONFIDENCE LEVEL ($\bar{X} + B$)
1	18	7.97	0.66	0.16	9.0	6.9	± 0.33	7.64	8.30
2, 3	36	7.54	0.65	0.11	9.0	6.3	± 0.22	7.32	7.76
7, 8	33	7.45	0.67	0.12	9.0	5.9	± 0.24	7.21	7.68
9	18	7.86	0.76	0.18	9.0	6.3	± 0.38	7.48	8.23
10, 11, 12	27	7.32	0.19	0.04	7.7	7.0	± 0.07	7.24	7.39
13, 14, 15	27	7.03	0.21	0.04	7.4	6.5	± 0.08	6.95	7.12
18, 19	20	6.67	0.32	0.07	7.2	6.0	± 0.15	6.52	6.82
20, 21	20	6.49	0.34	0.08	6.8	5.4	± 0.16	6.32	6.65
22, 23	20	6.23	0.27	0.06	6.7	5.8	± 0.13	6.10	6.35
24	24	7.48	0.44	0.09	8.4	6.8	± 0.19	7.29	7.67
25, 64	20	6.42	0.35	0.08	7.3	6.0	± 0.16	6.25	6.58
26	10	5.94	0.17	0.05	6.2	5.6	± 0.12	5.82	6.06
27	11	5.82	0.20	0.06	6.0	5.4	± 0.14	5.68	5.96
28, 29	18	5.63	0.27	0.06	6.1	5.2	± 0.13	5.50	5.77
30	9	5.53	0.19	0.07	5.9	5.3	± 0.15	5.38	5.78
33, 34	18	5.62	0.36	0.09	6.5	5.2	± 0.18	5.44	5.80
37	9	5.70	0.35	0.12	6.5	5.3	± 0.27	5.43	5.97
40	9	5.74	0.34	0.11	6.2	5.3	± 0.26	5.49	6.00
41	9	5.66	0.32	0.11	6.3	5.1	± 0.25	5.41	5.91
42	9	5.58	0.21	0.07	5.8	5.3	± 0.16	5.42	5.74
43	9	5.53	0.21	0.07	6.0	5.3	± 0.16	5.37	5.70
45	9	5.32	0.40	0.13	5.9	4.6	± 0.30	5.02	5.63
46, 47	34	4.82	0.29	0.05	5.4	4.3	± 0.10	4.70	4.92
48	17	4.89	0.35	0.09	5.6	4.5	± 0.18	4.71	5.08
49	17	4.91	0.32	0.08	5.4	4.5	± 0.16	4.74	5.07
50	17	4.47	0.38	0.09	5.2	3.9	± 0.20	4.28	4.67
51	17	4.23	0.30	0.07	4.9	3.8	± 0.15	4.08	4.38
52	17	4.17	0.31	0.07	4.9	3.6	± 0.16	4.01	4.32
53	17	4.11	0.25	0.06	4.5	3.7	± 0.13	3.98	4.23
54	17	4.21	0.26	0.06	4.6	3.5	± 0.13	4.08	3.34
55	17	4.22	0.24	0.06	4.7	3.8	± 0.13	4.09	4.34
56	17	4.19	0.17	0.04	4.5	3.9	± 0.09	4.10	4.28
57, 58	34	4.16	0.18	0.03	4.5	3.8	± 0.07	4.09	4.22
59, 60	34	4.64	0.32	0.06	5.4	4.2	± 0.11	4.52	4.75
61	17	5.04	0.41	0.10	5.7	4.4	± 0.21	4.83	5.25
62	17	5.49	0.28	0.07	5.8	4.8	± 0.15	5.35	5.64

TABLE 2

COMPARISON OF DATA - DISSOLVED OXYGEN

STREAM "A"

STATION 49 | STATION 50

TIME	D O PPM	TIME	D O PPM	TIME	D O PPM	TIME	D O PPM	D O PPM
SEPT. 27								
9.00 A.M.	8.2	3.00 P.M.	8.0	10.00 P.M.	7.3	4.00 P.M.	7.4	6.8
10.00 A.M.	8.0	4.00 A.M.	8.0	11.00 P.M.	7.0	5.00 P.M.	7.6	7.2
11.00 A.M.	7.8	5.00 A.M.	7.7	SEPT. 29		6.00 P.M.	7.4	7.4
12 NOON	7.6	6.00 A.M.	7.5	12 MIDNIGHT	7.3	7.00 P.M.	7.4	8.0
1.00 P.M.	7.8	7.00 A.M.	7.0	1.00 A.M.	7.5	8.00 P.M.	7.4	7.7
2.00 P.M.	7.8	8.00 A.M.	7.5	2.00 A.M.	7.1	9.00 P.M.	7.5	7.2
3.00 P.M.	7.8	9.00 A.M.	7.6	3.00 A.M.	7.3	10.00 P.M.	7.7	7.6
4.00 P.M.	7.6	10.00 A.M.	7.6	4.00 A.M.	7.5	11.00 P.M.	6.0	4.0
5.00 P.M.	7.6	11.00 A.M.	7.4	5.00 A.M.	7.5			
6.00 P.M.	7.8	12 NOON	8.4	6.00 A.M.	8.0	SEPT. 30		
7.00 P.M.	8.0	1.00 P.M.	8.0	7.00 A.M.	7.0	12 MIDNIGHT	7.7	
8.00 P.M.	8.0	2.00 P.M.	7.7	8.00 A.M.	7.5	1.00 A.M.	7.0	
9.00 P.M.	8.0	3.00 P.M.	7.8	9.00 A.M.	7.4	2.00 A.M.	7.0	
10.00 P.M.	8.0	4.00 P.M.	7.6	10.00 A.M.	7.6	3.00 A.M.	8.0	
11.00 P.M.	8.0	5.00 P.M.	8.2	11.00 A.M.	7.4	4.00 A.M.	7.5	
SEPT. 28		6.00 P.M.	8.0	12 NOON	7.8	5.00 A.M.	7.5	
12 MIDNIGHT	8.0	7.00 P.M.	7.6	1.00 P.M.	7.6	6.00 A.M.	7.7	
1.00 A.M.	8.0	8.00 P.M.	7.4	2.00 P.M.	7.8	7.00 A.M.	7.5	
2.00 A.M.	8.0	9.00 P.M.	7.3	3.00 P.M.	7.2	8.00 A.M.	7.7	

	STATION 49	STATION 50
MEAN	7.64	6.99
VARIANCE	0.1008	1.3911
STANDARD DEVIATION	0.3174	1.1794
F - TEST	$F\alpha$ CAL = 13.078 $F\alpha$ TAB = 4.00	α = 0.05
t - TEST	$t_{\alpha/2}$ CAL = 3.748 $t_{\alpha/2}$ TAB = 2.00	α = 0.05
CONF. FACT. $t_{\alpha/2,(n-1)} \frac{S}{\sqrt{n}}$	± 0.074	± 0.986
CONFIDENCE LIMITS	7.566 — 7.714	6.004 — 7.976

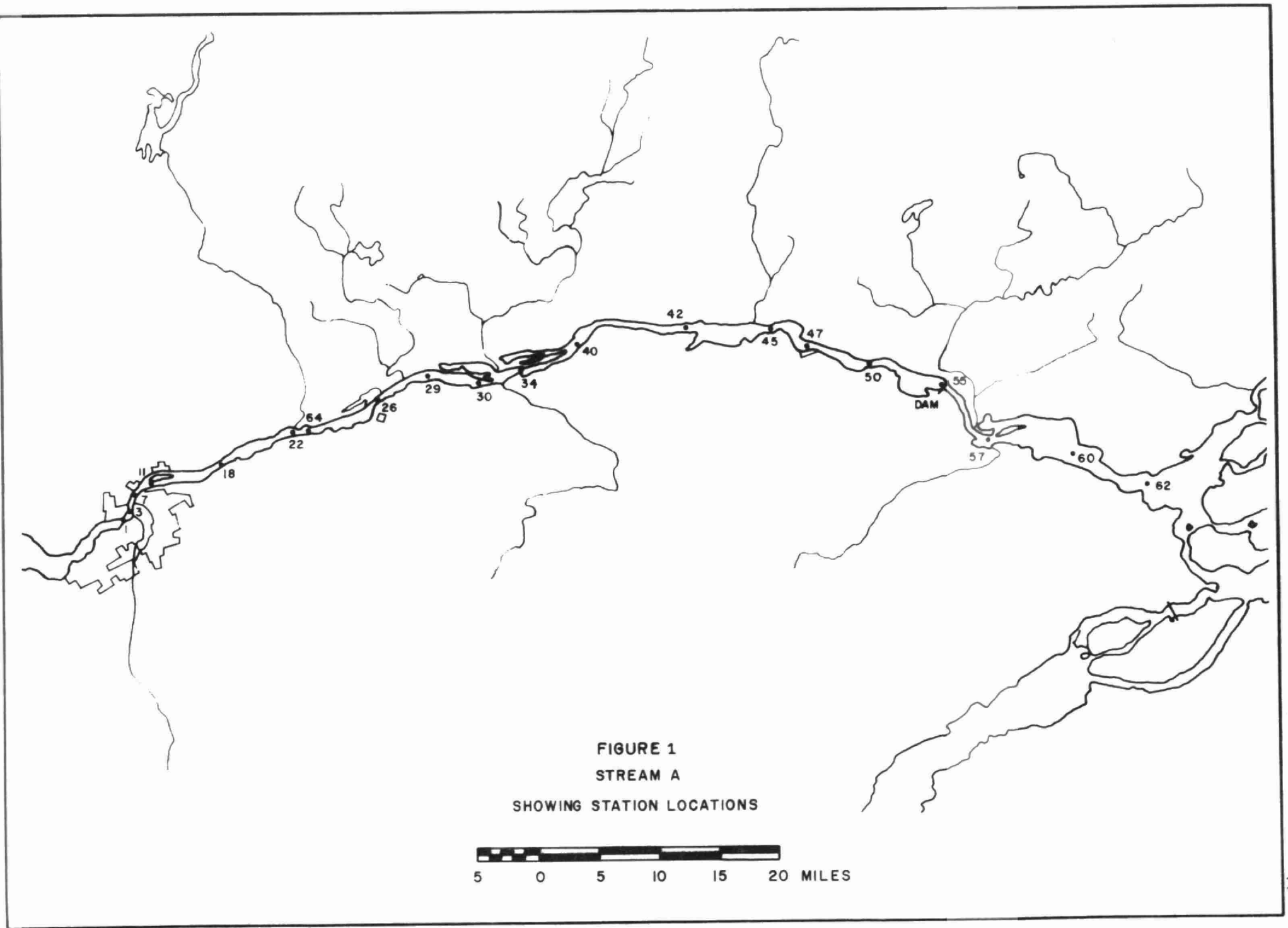
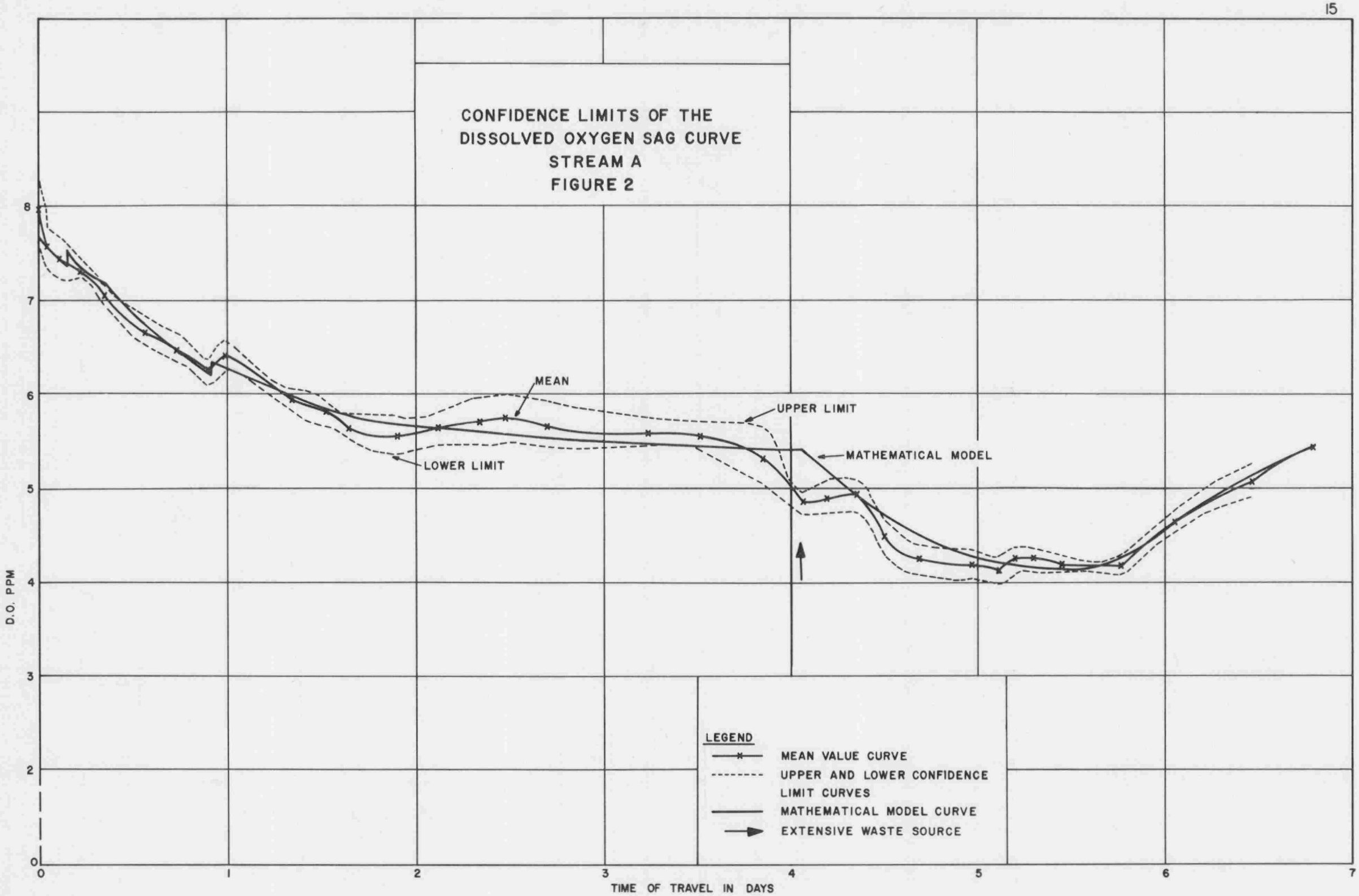


FIGURE 1

STREAM A

SHOWING STATION LOCATIONS

CONFIDENCE LIMITS OF THE
DISSOLVED OXYGEN SAG CURVE
STREAM A
FIGURE 2



APPENDIX - NOTATION

α	=	level of significance, probability of a Type I error;
$1 - \alpha$	=	degree of confidence;
BOD_5	=	5-day, 20°C., Biochemical Oxygen Demand;
DO	=	Dissolved Oxygen;
F	=	the F test: the parametric analysis of variance;
μ	=	the population mean;
n	=	the number of independently drawn cases in a single sample;
N	=	the number of cases in a population, the size of the population;
ppm	=	parts per million;
p(z)	=	probability associated with the value z;
s	=	sample standard deviation;
$s_{\bar{x}}$	=	$\frac{s}{\sqrt{n}}$ = standard deviation of the sample mean, or standard error of the sample mean;
σ	=	standard deviation of a population;
$\frac{\sigma}{\sqrt{n}}$	=	standard deviation of the population mean, or standard error of the mean;
t	=	Student's t test: a parametric test, the t distribution;
x	=	the random variable;

$\bar{x}$ = the sample mean;

z = $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$, the standardized mean;

$z_{\alpha/2}$ = the 100 ($\alpha/2$) percentage point of the normal distribution, here it is the area under the normal distribution curve from $z_{\alpha/2}$ to α so that it is equal to the value of $\alpha/2$.

REFERENCES

1. Freund, J. E., "Mathematical Statistics", Prentice-Hall, Inc., 1962. p.366, 203, 367.

BIBLIOGRAPHY

1. Neville, A. M., J. B. Kennedy, "Basic Statistical Methods for Engineers and Scientists", International Textbook Company, 1964.

2. Bowker, A. H., and G. J. Lieberman, "Engineering Statistics", Prentice-Hall, Inc., 1959.

3. Astin, A. V., "Experimental Statistics", National Bureau of Standards, Handbook 91, U. S. Government Printing Office, Washington, D. C.20402, August 1, 1963.